L'indice delle pagine dei doppi si rinnova con SWISH-E

Zeno Tajoli
CILEA, Segrate

Abstract
Il servizio "Scambi e doni tra Biblioteche" rinnova il suo indice full-text. Il sistema ora usato è l'opensource Swish-e. Vengono messe in evidenza i limiti del precedente sistema, le qualità del nuovo e le motivazioni della scelta. Lo sviluppo dell'interfaccia è stato calibrato sulla sua particolare utenza di bibliotecari. In fine vengono forniti degli esempi della configurazione adottata e spiegati gli interventi sul codice effettuati.

Keywords: Biblioteche, Indicizzazione pagine web, full-text, Swish-e, Virtual Library, doppi.

Il CILEA gestisce, all'interno del sito della "Virtual Library"[4], un elenco delle pagine con cui le biblioteche italiane segnalano i fascicoli doppi per scambiarli con altre biblioteche interessate a riceverli [2]. Oltre ai fascicoli di periodici, in numero minore, vengono segnalate anche delle monografie.
Parte delle pagine sono ospitate direttamente dal CILEA e vengono aggiornate ogni volta che la biblioteca invia degli aggiornamenti; la maggior parte però è ospitata sui siti istituzionali delle biblioteche. Si tratta sia di pagine dinamiche sia statiche o di documenti in formato Word, RTF, PDF. In particolare si è notato un aumento delle pagine dinamiche, che permettono un più facile e frequente aggiornamento da parte delle biblioteche.
In precedenza le pagine web erano indicizzate attraverso Index Server, un servizio di Microsoft IIS. Le pagine venivano copiate sul server del CILEA e poi indicizzate. Tramite appositi script in ASP venivano offerte diverse possibilità di ricerca e mantenuti i collegamenti alle pagine originali.
All'inizio del 2003 si è fatto il punto sulla situazione e si sono evidenziati questi problemi:

- la piattaforma in uso aveva problemi a gestire le pagine dinamiche;
- l'aggiornamento era problematico e per forza saltuario;
- poiché Index Server lavora in locale, le date d'aggiornamento che registrava non potevano essere quelle reali;
- il CILEA non sarebbe riuscito a gestire da solo i necessari miglioramenti allo strumento.

Pertanto a marzo 2003 è stata fatta un'indagine tra i diversi strumenti disponibili e si è identificato in SWISH-E[3] il più adatto alle esigenze del servizio.
In particolare questi punti sono stati per noi importanti:

- si tratta di un progetto abbastanza diffuso che diverse persone stanno portando avanti;
- i sorgenti sono completamente disponibili;
- le pagine dinamiche non hanno problemi ad essere indicizzate;
- è abbastanza semplice aggiungere un formato a quelli già supportati se si dispone di un generico reader per quel formato;
- è possibile usare le date effettive dei documenti indicizzati;
- l'aggiornamento dell'indice può essere automatizzato;
- si presta facilmente ad indicizzare singole pagine web che si trovano su diversi siti.

In particolare l'ultimo punto è stata la caratteristica fondamentale che lo ha fatto preferire ad altri strumenti come ad esempio ht://Dig o OpenFTS. Questi infatti sono sviluppati per

indicizzare un intero sito, non singole pagine web presenti su diversi siti.
La versione installata è la 2.2.3

L'interfaccia

Per sviluppare l'interfaccia si sono fatte le seguenti considerazioni.
Essa verrà usata essenzialmente da bibliotecari che vogliono completare le loro collezioni. Pertanto si ipotizza che conosceranno perfettamente o quasi il titolo da loro ricercato.
L'update delle pagine web, le fonti delle informazioni indicizzate, è ancora molto irregolare, dunque una pagina più recente è una fonte migliore di un pagina meno recente.
Per questi motivi l'interfaccia di default fa una "ricerca per frase", cioè considera quanto immesso come il titolo di un periodico. Non si richiede che il titolo sia immesso completamente.
Gli ordinamenti e i filtri previsti sono solo per data.
Come opzione aggiuntiva è fornita la ricerca per parole chiave con gli eventuali operatori booleani. L'help è scarno proprio perché presuppone una certa competenza nell'uso di strumenti di ricerca.

Configurazione e modifiche effettuate al software

L'installazione e la configurazione di base sono molto ben documentati sul sito di riferimento[3] (v. anche l'articolo[1]). Il software è composto da un programma in C che crea gli indici e governa le ricerche, da dei file di configurazione, infine da programmi in Perl che gestiscono la conversione dei formati e l'interfaccia HTML
I file di configurazione di base sono swish.conf e spider.conf. Per indicizzare un numero non elevatissimo di singole pagine web è stata fatta questa configurazione:
swish.conf

```
# Program to read documents
IndexDir ./spider.pl
# Define the config file for the spider to use
SwishProgParameters spider.conf
# Use libxm2 for parsing documents
DefaultContents HTML*
IndexContents TXT* .txt .text
# Cache document contents in the index for context
display
StoreDescription HTML <body>
StoreDescription HTML2 <body>
```

spider.conf

```
my %Server1 = (
base_url => 'http://www.agr.unipi.it/biblio/doppi.htm',
email        => 'tajoli@cilea.it',
delay_min     => .2,
max_size      => 1_000_000,
max_depth     => 0,
keep_alive    => 1,
test_url      => \&test_url,
);
[...]
my %ServerXXX = (
base_url => 'http://meneghetti.univr.it/CATDOP99.doc',
email        => 'tajoli@cilea.it',
delay_min     => .2,
max_size      => 1_000_000,
max_depth     => 0,
keep_alive    => 1,
test_url      => \&test_url,
filter_content => \&doc,
);
@servers = (\%Server1, ServerXXX, );
sub test_url {
  my ( $uri, $server ) = @_;
  return $uri->path;
}
use pdf2html;
sub pdf {
my ( $uri, $server, $response, $content_ref ) = @_;
return 1 unless $response->content_type eq
'application/pdf';
$server->{counts}{'PDF transformed'}++;
$$content_ref = ${pdf2html( $content_ref, 'title' )};
$$content_ref =~ tr/ / /s;
return 1;
}
use doc2txt;
sub doc {
my ( $uri, $server, $response, $content_ref ) = @_;
return 1 unless $response->content_type eq
'application/msword';
$server->{counts}{'DOC transformed'}++;
$$content_ref = ${doc2txt( $content_ref )};
$$content_ref =~ tr/ / /s;
return 1;
}
 1;
```

Le modifiche hanno interessato i programmi in Perl.
La parte maggiormente modificata è stato il modulo TemplateDefault.pm; si tratta del modulo che gestisce la creazione dell'interfaccia HTML, gli interventi sono stati: traduzione di etichette, aggiunta di una nuova opzione di ricerca cioè la scelta della ricerca per frase o per parola chiave, modifica della sintassi HTML per renderla HTML 4.01 compatibile.
Solo per esigenze di traduzione si è intervenuto sul modulo DateRanges.pm.
Gli interventi più complessi sono stati quelli sulla CGI perl swish.cgi.

Sono stati effetuati sia interventi di traduzione e di configurazione sia l'aggiunta del codice necessario a gestire la nuova opzione "Ricerca per frase / Ricerca per parole chiave".
In generale l'interfaccia è stata tradotta in italiano ma non completamente; le segnalazioni di errore sono rimaste in inglese perché Swish-E non le gestisce con dei templates ma con messaggi inseriti nel codice perl.

Riferimenti

[1] Rabinowitz Josh, "How to Index Anything", Linux Journal, n. 111, July 2003. URL: http://www.linuxjournal.com/article.php?sid=6652
[2] Scambi e doni tra biblioteche. URL: http://www.cilea.it/Virtual_Library/bibliot/doppi.htm
[3] SWISH-Enhanced. URL: http://swish-e.org/
[4] Virtual Library CILEA.. URL: http://www.cilea.it/Virtual_Library/